Repeatability & Workability Evaluation of SIGMOD 2009*

S. Manegold[1] I. Manolescu[2] L. Afanasiev[3] J. Feng[4] G. Gou[5]
M. Hadjieleftheriou[6] S. Harizopoulos[7] P. Kalnis[8] K. Karanasos[2] D. Laurent[9]
M. Lupu[10] N. Onose[11] C. Ré[12] V. Sans[9] P. Senellart[13] T. Wu[14]
D. Shasha[15]

[1] CWI, Netherlands
[2] INRIA Saclay–Île-de-France, France
[3] University of Amsterdan, Netherlands
[4] Sun Yat-Sen University, China
[5] Microsoft Corporation, USA
[6] AT&T Labs - Research, USA
[7] HP Labs, USA
[8] KAUST, Saudi Arabia
[9] ETIS, Univ. de Cergy-Pontoise, France
[10] Information Retrieval Facility, Vienna, Austria
[11] University of California, Irvine, USA
[12] University of Wisconsin, Madison
[13] Télécom Paristech, France
[14] University of Illinois at Urbana-Champaign, USA
[15] Courant Institute, New York, USA

ABSTRACT

SIGMOD 2008 was the first database conference that offered to test submitters' programs against their data to verify the repeatability of the experiments published [1]. Given the positive feedback concerning the SIGMOD 2008 repeatability initiative, SIGMOD 2009 modified and expanded the initiative with a workability assessment.

1. THE GOAL

On a voluntary basis, authors of accepted SIGMOD 2009 papers provided their code/binaries, experimental setups and data to be tested for:

repeatability of the experiments described in each accepted paper;

workability of the software by running different/more experiments with different/more parameters than shown in the accepted paper;

by a repeatability/workability committee (which we call the *RWC*), under the responsibility of the repeatability/workability editors-in-chief (which we call the *RWE*).

2. THE PEOPLE

The RWE were Ioana Manolescu and Stefan Manegold. The 2009 RWC consisted of the other authors of this paper, along with D. Shasha.

*`http://homepages.cwi.nl/~manegold/SIGMOD-2009-RWE/`

3. THE PLAN

Several lessons learned from the first repeatability evaluation with SIGMOD 2008 [1] led us to improve and extend the process. The following paragraphs describe the details.

3.1 Accepted papers, only

The SIGMOD 2009 repeatability & workability committee evaluated accepted papers only. The primary reason for this change was to reduce the workload for the evaluation by avoiding evaluation of papers that would eventually not be accepted. A second reason was that authors had commented that they wouldn't mind the extra work of preparing their repeatability & workability submission once their papers were accepted.

3.2 Adapted schedule

Focussing on accepted papers only required an adaptation of the general schedule for the repeatability & workability evaluation. After the SIGMOD 2009 program committee had announced the accepted papers, the contact authors of all accepted research papers were personally invited via email to prepare and submit their experiments including code, data sets and detailed instructions. This later start of the evaluation did not leave enough time to finish the evaluation before the camera ready deadline, thus preventing authors from mentioning the result of the evaluation in the final versions of their papers. In fact, the evaluation was completed just before the conference to give the authors the chance to mention the results in their presentations at SIGMOD 2009.

3.3 Refined submission method

In contrast to the push-based submission in 2008 via upload to a FTP server, the submission in 2009 was pull-based. Authors were asked to make their submissions available for download by the RWE. This helped to avoid problems with uploading large (sometimes tens of gigabytes) submissions to a single FTP server. The RWE then made the submissions available for download to the assigned reviewers.

3.4 Refined submission instructions

To give the reviewers some information to better plan their evaluation, the authors were asked to include in their submission information the length of time their experiments were expected to run. In addition, in order to facilitate the workability evaluation, the authors were asked to extend their repeatability instructions with suggestions as to how to extend their experiments beyond the contents of their paper. Possibilities ranged from explanations of how to use different data sets, query work loads, tuning and/or configuration parameters to compilation, and installation instructions for alternative hardware/software environments.

3.5 Refined reviewing process

As in 2008, the assignment of papers to reviewers was mainly determined by the need to match the papers' hardware and software requirements with the reviewers resources. Of course, (potential) conflicts of interest were avoided. In contrast to 2008, each paper was assigned two reviewers: a *primary* reviewer to do the actual repeatability and workability evaluation, and a *secondary* reviewer as back-up and to double-check the primary reviewers report.

3.6 Author-reviewer-interaction

The 2008 experiences revealed that successful repeatability evaluation can be hindered or even prevented by minor problems in setting up and running the experiments due to missing details in the provided instructions. To solve this problem, the 2009 effort provided a web-based anonymous communication channel to allow interaction between authors and reviewers to resolve problems as early as possible. All communication has been archived. With standard WIKI or BLOG software either not providing convenient and effective means for anonymous peer-to-peer communication, we wrote a PHP script to efficiently provide the basic functionality required.

4. THE PROCESS

After the announcement of the accepted papers, the contact authors of all 64 accepted research papers were invited by email to prepare and submit their contribution. By the (extended) deadline of April 22 2009, 19 authors had provided their contribution. The remaining 45 authors chose not to reply at all. In contrast to 2008, authors were not asked to provide an explanation why they could not submit their code, data and experiments for evaluation.

Each RWC member was assigned three papers, either two for primary review and one for secondary review, or one for primary and two for secondary review. Assigned reviewers met the software and hardware requirements of the experiments though sometimes at significant effort. For example, some reviewers installed extra software or even (re-)install complete machines. In one case, the reviewer's group (re-)installed a 40-node Linux PC cluster to repeat a scaled-down version of experiments that were originally run on a 100-node cluster.

In nearly all cases, the anonymous web-based communication channel between authors and reviewers was successfully used to resolve problems ranging from missing gnuplot files to insufficient specification of the versions of required software. In only two cases was the discussion insufficient to solve all problems, resulting in only a partial repeatability evaluation for those papers.

The reviewing process stretched over the complete two month period, with the final reviews being finished only the day before SIGMOD 2009 started. The long time partly due to (i) the installation of extra hardware and software, as well as configuration work required; (ii) author-reviewer communication to solve initial problems; and (iii) experiments that took several days or even weeks to run.

Not all authors provided hints how to modify and/or extend their experiments for workability evaluation. In all but 5 cases, the reviewers managed to find their own ways to modify/extend the respective experiment to assess their workability. Even when workability suggestions were provided, the reviewers volunteered to go beyond the authors' suggestions.

5. THE RESULTS

Overall, the results of the evaluation for the 19 submissions were rather positive:

- For 10 papers, the presented experiments could be fully repeated and workability was confirmed.
- For 1 paper, repeatability was fully confirmed and workability was mostly confirmed.
- For 4 papers, all original experiments were successfully repeated, but workability was not, mostly due to missing or insufficient instructions on how to modify the original setup conveniently.
- For 1 paper, the experiments were mostly repeated, but workability could not be evaluated.
- For 1 paper, the repeatability evaluation was successful, but the workability evaluation failed.
- For 2 papers, major technical problems could not be solved within the two months reviewing period, preventing most or all of the repeatability and workability evaluation.

The authors were informed before the conference about the results for their papers, and thus given the opportunity to mention the results during their presentation at SIGMOD 2009.

6. THE ASSESSMENT

With many of the lessons learned from the 2008 effort, the 2009 repeatability and workability evaluation went much smoother than the previous round. In particular focussing on accepted papers only (an idea suggested by Donald Kossmann), pull-based submission, and the web-mediated discussions between reviewers and authors to solve minor technical problems proved to be successful.

Though creating a higher workload for the reviewers, the newly introduced workability evaluation (when successful) gave even more credibility to the authors than a pure repeatability evaluation.

The unexpectedly low submission rate appears to be due to the fact that the authors were not aware of the SIGMOD 2009 repeatability & workability evaluation by the paper submission deadline. Due to several delays and issues, the SIGMOD 2009 repeatability & workability evaluation was not announced in any call for papers, nor mentioned

on the SIGMOD 2009 web site. Several personal communications with authors during the conference revealed that many authors were caught by surprise when invited to submit repeatability material for their accepted papers, or were simply not sure how "official" the evaluation was. In other words, there was probably insufficient publicity around the SIGMOD 2009 repeatability & workability evaluation.

While serving its primary purpose, the PHP script for the reviewer-author-communication could be improved. Not being a standard tool, the "look-and-feel" was considered "unusual" and the automatic email notification of new postings did not always work reliably.

Given the diversity of the papers and their experiments, the reviewers were not given a strict format for their reviews, but rather allowed to freely determine the format, structure and content of their reviews themselves, to accommodate the process they followed as well as their findings and final verdict.

7. RECOMMENDATIONS

Here are some lessons for 2010 and beyond:

- Publicize the effort well in advance of the submission deadline.
- Improve the author instructions, in particular to ask more explicitly for workability instructions. More generally, collecting, improving and disseminating guidelines for the preparation of repeatable experiments requires more work in the community; tutorials such as [2] are a step in this direction.
- Improve the reviewer guidelines to unify the results. Given the diversity of the experiments, this is not a trivial task, as any guidelines and/or format still need to leave sufficient room for all cases.
- Improve the visibility of the evaluation results. The SIGMOD PubZone server [3] is a promising tool for this purpose.
- Improve the software support for author-reviewer discussions. With respect to the last item above, we are currently considering the extension of the MyReview conference management tool [5] to accomodate the specific needs of our process. The main feature we need, and which is not yet supported by MyReview and other similar tools, is the possibility for reviewers and authors to exchange an unbounded number of messages, over the whole period of reviewing (as opposed to one single exchange, at a specific point in the process, as currently used for conferences such as ACM SIGMOD, and supported by the Microsoft Research Conference Management Tool [4]).

8. SUMMARY

Our community can be justly proud of its practical impact over the years. This is due to a confluence of good ideas with good engineering. Repeatability and Workability help to ensure the validity of good ideas and provide paradigms and platforms for good engineering.

Appendix

The following SIGMOD 2009 accepted papers passed all the repeatability tests, and also passed workability test:

Authenticated Join Processing in Outsourced Databases by Yin Yang, Dimitris Papadias, Stavros Papadopoulos and Panos Kalnis

Scalable Join Processing on Very Large RDF Graphs by Thomas Neumann and Gerhard Weikum

Self-organizing Tuple Reconstruction in Column-stores by Stratos Idreos, Martin Kersten and Stefan Manegold

A Revised R-tree in Comparison with Related Index Structures* by Norbert Beckmann and Bernhard Seeger

An Architecture for Recycling Intermediates in a Column-store by Milena Ivanova, Martin Kersten, Niels Nes and Romulo Goncalves

Skip-and-Prune: Cosine-based Top-K Query Processing for Efficient Context-Sensitive Document Retrieval by Jong Wook Kim and K. Selcuk Candan

Incremental Maintenance of Length Normalized Indexes for Approximate String Matching by Marios Hadjieleftheriou, Nick Koudas and Divesh Srivastava

Cost Based Plan Selection for XPath by Haris Georgiadis, Minas Charalambides and Vasilis Vassalos

Core Schema Mappings by Giansalvatore Mecca, Paolo Papotti and Salvatore Raunich

Secondary-Storage Confidence Computation for Conjunctive Queries with Inequalities by Dan Olteanu and Jiewen Huang

Minimizing the Communication Cost for Continuous Skyline Maintenance by Zhenjie Zhang, Reynold Cheng, Dimitris Papadias and Anthony K. H. Tung

The following SIGMOD 2009 accepted papers passed repeatability tests:

A Comparison of Approaches to Large Scale Data Analysis by Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden and Michael Stonebraker

Query Simplification: Graceful Degradation for Join-Order Optimization by Thomas Neumann

ROX: Run-time Optimization of XQueries by Riham Abdel Kader, Peter Boncz, Stefan Manegold and Maurice van Keulen

Simplifying XML Schema: Effortless Handling of Nondeterministic Regular Expressions by Geert Jan Bex, Wouter Gelade, Wim Martens and Frank Neven

Secure k-NN Computation on Encrypted Databases by Wai Kit Wong, David Wai-lok Cheung, Ben Kao and Nikos Mamoulis

Detecting and Resolving Unsound Workflow Views for Correct Provenance Analysis by Peng Sun, Ziyang Liu, Susan B. Davidson and Yi Chen

Acknowledgments We are thankful to a set of external referees: Tuyet Tram Dang Ngoc, Tao-Yuan Jen and Dan Vodislav from ETIS - Université de Cergy-Pontoise; Guangxishui Yang, Yi Zhou, and Junyu Liu from Sun Yat-Sen University; Jeffrey Xu Yu, Lijun Chang and Lu Qin from the Chinese University of Hong Kong; and YongChul Kwon from the University of Washington, Seattle.

9. REFERENCES

[1] I. Manolescu, L. Afanasiev, A. Arion, J.-P. Dittrich, S. Manegold, N. Polyzotis, K. Schnaitter, P. Senellart, S. Zoupanos, and D. Shasha. The repeatability experiment of SIGMOD 2008. *SIGMOD Record*, 37(1):39–45, Mar. 2008.

[2] I. Manolescu and S. Manegold. Performance Evaluation in Database Research: Principles and Experience. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, Cancún, Mexico, 2008. Tutorial slides are available from `http://www.icde2008.org/` or from the authors. A shortened version was presented also at the EDBT 2009 conference.

[3] PubZone: scientific publication discussion forum. `http://www.pubzone.org`.

[4] The Microsoft Research Conference Management Tool. `https://cmt.research.microsoft.com`.

[5] The MyReview Conference Management System. `http://myreview.lri.fr`.